Of Gods and Goats: Weakly Supervised Learning of Figurative Art

Of Gods and Goats: Weakly Supervised Learning of Figurative Art

Elliot J. Crowley
elliot@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Department of Engineering Science
University of Oxford

Abstract

The objective of this paper is to automatically annotate images of gods and animals in decorations on classical Greek vases. Such images often require expert knowledge in labelling. We start from a large dataset of images of vases with associated brief text descriptions.

We develop a weakly supervised learning approach to solve the correspondence problem between the descriptions and unknown image regions. The approach progressively strengthens the supervision so that eventually a Deformable Part Model (DPM) sliding window detector can be learnt (for each god/animal) and used to annotate all vases by detection. There are two key steps: first, text mining the vase descriptions to obtain clusters for each god where there is a visual consistency between at least a subset of the images; and second, discriminatively searching within these clusters for consistent regions that are used as positive training examples for the DPM.

The method successfully annotates a large variety of Gods and other animals, and we include a quantitative evaluation over hundreds of images.

1 Introduction

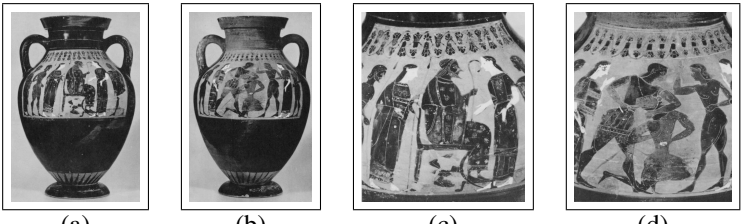
A number of papers have explored the "words and pictures" problem of automatically annotating image regions with particular words, given only images and associated text [3, 4, 5, 6, 11]. For example, in "Names and Faces in the News" [6] the problem is to label the faces in images accompanying stories on news web pages, and this leads to a correspondence problem because there may be several faces in each image and several people named in the news story. In general, the problem requires discovering co-occurrences between image regions and words over a large set of paired image-words data, and consequently the algorithms employed have used ideas from machine translation [21, 25] and weakly supervised learning [8, 14].

Our goal in this paper is to automatically annotate the decorations on classical Greek vases with the gods and animals depicted, given a large dataset of images of vases with associated short text descriptions. It thus falls into the problem area of "words and pictures" but with the additional challenges of: (i) quite noisy supervision – only a subset of the images associated with the vase may show the scene described; (ii) non-naturalistic renderings – these are not images of real scenes, such as faces, but stylized figures in binary tones for which standard visual descriptors may not be suitable; (iii) most images contain multiple

figures with only quite subtle differences in appearance; and (iv) each god is depicted in a number of different styles/poses depending on the story being illustrated.

Figure 1 illustrates the small inter-class variability between gods, and Figure 2 the high intra-class variability for Zeus even within a distinctive pose (two poses are shown: seated and chasing unfortunate maidens).

To solve this problem we propose a weakly supervised learning approach that proceeds in a number of stages. The key idea is to use each stage to strengthen the supervisory information available, so that learning can finally proceed successfully – this is like improving the signal (supervision) to noise (irrelevant images, irrelevant image regions) in signal processing. The first stage (section 3) uses text mining methods to select sets of images that are visually consistent for a god depicted in a particular style. The second stage (section 4) then reduces the search space by eliminating irrelevant image regions. At this point the weak supervision is sufficiently strong, that a form of multiple instance learning [26, 27] can be used to identify the image regions depicting the god in those images where he/she appears. Finally, (section 5) the image regions are used to train a Deformable Parts Model (DPM) sliding window classifier [13] and all images in the dataset associated with the god can then be annotated by object category detection.



(a) (b) (c) (d)

A. Theseus and minotaur, with youths and women; B. Zeus seated, between women and onlookers;

(a) (b) (c) (d)

A. Dionysos with vine and drinking horn between satyr and maenad with oinochoe;
B. Apollo, playing kithara, between Artemis and leto, deer;

Figure 1: **Two example vase entries** each consisting of four images and two scene descriptions (given under each set). Note that in both examples the images have been ordered so that description A refers to images (b) & (d), and B to (a) & (c). Image (c) is a zoomed detail of (a), and (d) a zoomed detail of (b). However, the actual images are unordered, and the correspondence between description and images is unknown, as is the geometric relation between images.

2 Data – the Beazley Vase Archive

The publicly available pottery database of the Beazley Archive [1] contains around 50,000 vase entries with one or more associated images (for a total of 120,000 images). Each entry describes a particular vase; providing detailed information such as its date of origin, shape, painting technique, as well as a description of the scenes depicted on the vase. Figure 1 shows two typical entries.

Subject	No. Vases	No. Images	No. Clusters	Average No.	Total No.
Apollo	1598	3452	16	151	2417
Artemis	676	1676	14	37	521
Athena	3759	7457	14	162	2267
Dionysos	5533	9061	22	295	6491
Herakles	4045	7675	16	179	2858
Hermes	2206	5141	8	72	573
Poseidon	450	1276	12	32	384
Zeus	615	1709	4	176	705
God Total	18882	37447	106	138	16216

Table 1: **Subject totals over the dataset**. From left to right: the number of vases associated with each subject, the number of images associated with those vases, the number of visually consistent clusters (as described in section 3), the average number of images within each cluster and the total number of images across the clusters. Note that less than 50% of the images associated with vases for a subject will actually contain that subject. Also note that a vase can have multiple subjects so vases and images will overlap between gods. There are 727 vases with goats.

Each vase entry can have a number of associated *subjects* such as prominent gods or characters of note. These subjects feature in at least one of the scene descriptions. For the majority of vases there are two scene descriptions, one for each of the front and back sides, and a set of images associated with each side. In less than 8% of cases the subject will appear in multiple descriptions. The number of vases assigned to a particular subject and the total number of images belonging to those vases is given in table 1.

Geometric Relations and Distinguished Images. There is a large amount of redundancy in the images of this database. For example, in figure 1 there are two disjoint subsets of images $\{(a),(c)\}$ and $\{(b),(d)\}$, corresponding to the front and back of the vase, and within each disjoint subset one image is a zoomed detail of the other. In this case we only need to retain the two zoomed details (c) & (d) as these have the highest resolution and the remaining images are not required for learning the annotation. Note, small details are lost for the characters at the edge, but such characters are rarely the focal point of a scene. We refer to the subset of images that are retained for each vase set as the *distinguished images*.

To obtain the geometric relations affine transformations [18] are computed automatically between all pairs of images of the vase set. A disjoint subset is obtained by selecting images that are mutually related to each other, but not to other images. For example, images $\{A,B,C\}$ form a disjoint subset if there exists an affine transformation between A & B, B & C, and A & C, but not to any other image in the vase set. The most zoomed image from each disjoint subset is selected as the distinguished image.

3 Text mining for visually consistent clusters

The goal of this section is to select for each god several visually consistent clusters of vases that depict the god in a single pose. For example, for Zeus, clusters include those that depict the 'seated' pose, and those that depict the 'pursuing' pose as in figure 2. This clustering is required as otherwise there is too much variation in the images associated with a god to find out which elements are in common.

The visually consistent clusters for a particular god are obtained by mining the text descriptors for distinctive keywords, and then selecting all the vases containing those keywords. Keywords that correlate strongly with visual consistency are verbs (e.g. sitting, struggle, fighting, playing), and nouns such as animals and objects (e.g. lion, mule, lyre, bow), but not another person (human or god). The position of the keywords is also important – the

keyword should be after the god's name but before a word corresponding to a person. Determining such keywords and their position relative to the subject is the core of successfully mining for visually consistent clusters. The importance of keywords and their placement has been noted previously [5, 12, 16, 19].

Two examples of visually consistent clusters for the god Zeus are shown in figure 2, and table 1 gives the number of clusters mined for each god. Note, animals do not typically exhibit pose variation and may each be represented by a single visually consistent cluster.

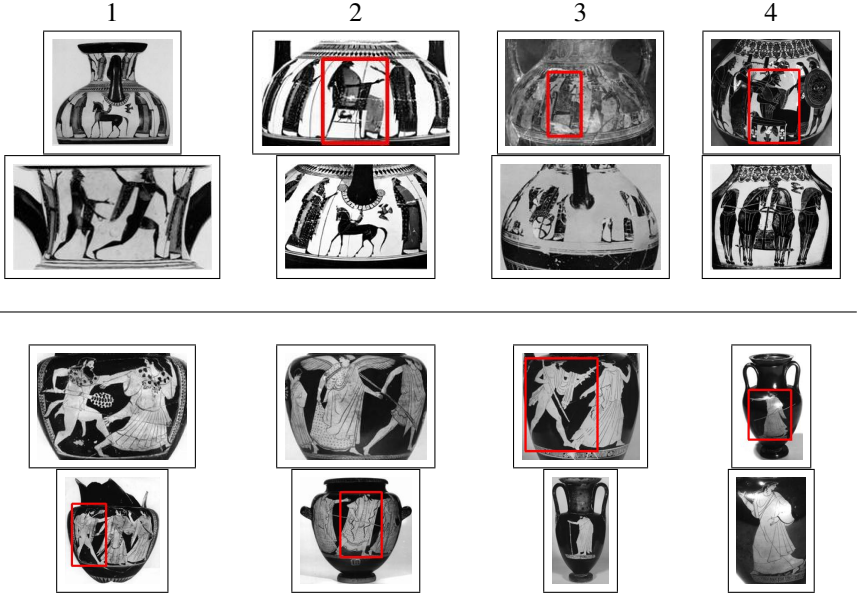


Figure 2: **Visually consistent clusters.** A subset of the vases in a visually consistent cluster. Top half: 'Zeus Seated' (4 vases shown from a cluster of 170 vases). Bottom half: 'Zeus Pursuing' (4 vases from a cluster of 94 vases). In both cases Zeus, where present, is indicated by a red rectangle. Note, the data is noisy: if perfect, Zeus would be expected to be in one distinguished image for each vase, but sometimes he is not; this is because for some vases there are text descriptions with no associated images.

Implementation details. During text mining, any stop-words (i.e. the, and, ...) are ignored. Clusters are obtained from all the vases associated with the particular god. For these vases only the scene description that contains the subject's name is used, and all words that occur after the subject but before the following person are extracted and aggregated. The extracted words are then ordered from most to least frequent, and the keywords are selected as the highest ranked words in this list. A greedy approach is then used to form the clusters: starting with the most frequent keyword, those vases containing the keyword are selected to form the first visually consistent cluster and removed from further consideration; then the second most frequent keyword is selected and so on. Table 2 illustrates this process for Herakles.

There are two main techniques of illustration in the Beazley Archive: black figures on red backgrounds (black-figure) and red figures on black backgrounds (red-figure). Images of these techniques are kept separate leading to two visually consistent clusters for each keyword.

We did investigate other methods of clustering: K-means clustering of the text descriptions and images (using visual words [30]) produced groupings with little visual consistency.

Keyword	lion	bull	club	suspended	quiver	bow	shield	boar
# occur.	884	309	259	232	209	194	174	137
# vases	884	300	123	49	11	36	141	90
# images	1560	336	245	108	13	78	345	173

Table 2: **Keyword mining for Herakles**. The most commonly occurring keywords (top row) with their corresponding frequency (second row). The third and fourth rows are the number of vases and distinguished images assigned to each visually consistent cluster by the greedy algorithm. Since vases are extracted in order starting from the highest ranked keyword (lion), some words (e.g. suspended) have far fewer vases than word occurrences.

We also investigated using the *Apriori* Algorithm [2] to find groups of words that occur together often. However, these groups often describe exact scenes (e.g. Judgement of Paris) and they were found: (i) to be too specific leading to small clusters, and (ii) not to correspond to a consistent pose. As will be seen in the sequel the proposed technique is very successful in generating visually consistent clusters with high recall and precision.

4 Searching for Candidate Regions

The aim of this section is to find the regions within the images of a visually consistent cluster that correspond to the subject of the cluster (i.e. a god in a particular pose or an animal). These windows are then used in section 5 to train a strong object category detector using a DPM [13], and this is used to find the subject across the whole database.

The task is the following: each vase in the cluster is represented by a set of distinguished images, and one of these images may contain the subject. However, which of the distinguished images to choose and also the region within this image that contains the subject is unknown.

We find the images and region using a discriminative approach based on multiple-instance-learning [27]. There are three stages: First, a reduced search area is obtained for each image (compared to the entire image) by eliminating regions that occur frequently in vases outside the cluster; second, multiple instance learning is used to discriminatively find candidate regions within the reduced search area that occur over multiple vases within the cluster but not in vases outside the cluster; third, since some of the candidates may not contain the subject, a final round of voting is used to select the veridical visually consistent regions and determine their spatial extent. These steps are next described in more detail.

1. Reducing the search space. Regions of the vase that repeat throughout the dataset, such as decorative patterns, can be excluded when searching for the god-region. To achieve this we employ the method of Knopp *et al.* [22] who removed confusing regions in images of street scenes. Figure 3 illustrates the type of regions that can be excised by this process. In addition to removing the commonly occurring regions, the search area is also restricted to the vase itself by identifying the background. This is achieved by segmentation using a modified version of GrabCut [23, 28].

2. Finding candidate windows for the gods. The key idea here is to propose candidate regions that may contain the god, and then test the hypothesis by determining if the region occurs in other images of the cluster, but not in images not containing the god as subject. In more detail, the candidate region is used to train a HOG [9] sliding window classifier (in the manner of an exemplar SVM [24] using one positive sample) but implemented here using the more efficient LDA training of [17] which does not require mining for hard negatives. The sliding window detector is applied to all images in the cluster and also to an equal number of images not containing the subject that are chosen randomly (in fact two detectors are

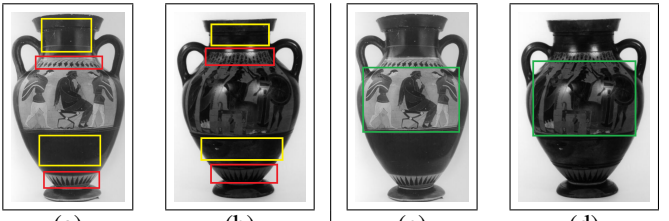


Figure 3: **Reducing the spatial search space**. The red rectangles in (a) & (b) delimit regions containing patterns commonly repeating across the dataset. The yellow rectangles delimit low gradient regions. They are extracted automatically as described in section 4. These regions are excluded when searching for the god-region. (c) & (d) show the corresponding search regions of (a) & (b).

learnt for every candidate to account for left-right flipping of the pose). In the language of multiple-instance-learning, the distinguished images of each vase are the positive-bags (and the detector should select an instance on one of these) and the images not containing the subject form the negative bag. Note, as shown in figure 2 some of the 'positive' bags do not contain the god. Furthermore, the instance can be anywhere within the reduced search region on each image.

Candidate regions are proposed on all distinguished images with at least $2\times$ zoom relative to another image in their disjoint subset (see section 2). The candidates are sampled around the centre of each image as this is where the subject of interest tends to occur. The windows are made large enough to be discriminative but not large enough to incorporate too much of the scene beyond the subject. Two partly overlapping windows are sampled on each image. Candidate windows with low gradient energy are rejected as being uninformative.

For each candidate window, all images are ranked by the highest detector score, and the candidate is scored by the number of detections made on the positive bags before any detections are made on the negative bag. The 20 candidates that score highest in this way are kept and the remainder discarded. If less candidates are used high-scoring mistakes will have too much influence on the outcome, if more candidates are used there are simply more mistakes present. This method of discriminatively finding regions is somewhat similar to the weak learning method of Singh *et al.* [29] and Juneja *et al.* [21]

3. Obtaining visual consistency. The candidates regions have been extracted independently and we now seek visual consistency amongst the 20 candidates to ensure that the god region is found, and eliminate outliers in the candidates. To this end, each of the candidate regions 'votes' for the others, and the candidates with the most votes are retained. In detail, the LDA detector for each candidate is applied to the remaining 19 images (where the other candidates have been learnt from). The top detection made on each image is compared to the candidate window associated with that image and if the overlap ratio (intersection / union) exceeds 0.5 then the candidate window on that image obtains one vote. Note that multiple windows from the same image can be present in the top 20 windows; it is the windows that are considered, not the images.

The four windows with most votes are retained and for these the window size is re-estimated in a robust manner using the median of its candidate window and the overlapping detections made by the other remaining windows. This re-estimation improves the god-coverage where the original candidates were not well aligned. We reduce the number of windows to four to reduce the likelihood that there are false positives present. The resulting windows for two different visually consistent clusters are shown in figure 4.

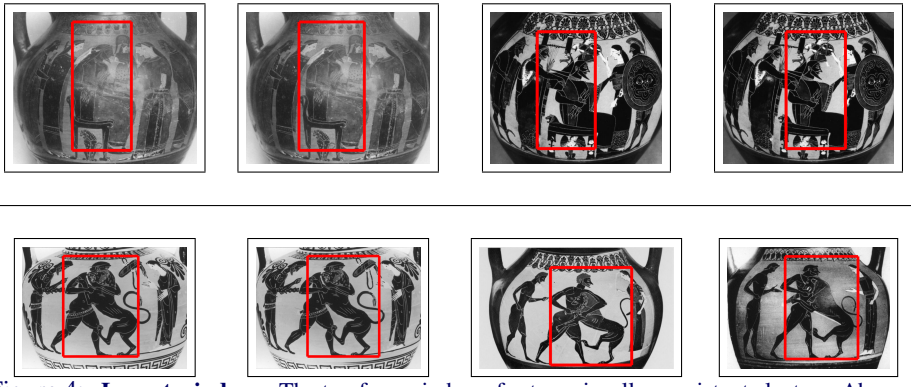


Figure 4: **Learnt windows**. The top four windows for two visually consistent clusters. Above: Zeus Seated. Below: Herakles Lion.

Discussion. There are a number of existing methods [7, 10, 31] for isolating the regions within images containing an object, given only image level annotation. However these tend to require that the object be in each image considered and are as such not appropriate for our noisy annotation task. A recent method [15] could provide an alternative method of achieving visual consistency amongst the candidates.

5 Training Strong Detectors

Up to this point we have been restricting the set of searchable images for each subject to those found in visually consistent clusters. The weak supervision is now strong enough to train strong detectors that can be run on a larger set of images. The training proceeds in two stages: first a LDA model generates additional positive examples within the visually consistent cluster, then these examples are used to train a stronger DPM [13].

Obtaining further positive examples. For each cluster there are four windows that best represent the visually consistent aspect of the corresponding subject. These four windows are averaged to form the mean input for a LDA detector. This is then applied to all other distinguished images in the cluster, and the the top scoring detection made on each image is considered. These are ordered by classifier score. We choose which detections to use as further examples using an adaptive threshold: recall that each vase is represented by several distinguished images, typically only one of which will contain the subject. This means that only one firing is likely correct on a vase. When another firing is made on that vase it is a error. We accept each detection in order as a positive window until two errors have been made. We choose to allow for two errors as there is a small chance a subject will appear twice on a vase.

Training a DPM. For each visually consistent cluster, the four windows and the additional windows obtained above are used as positive training examples for a DPM consisting of one component and four parts. Negative examples are chosen at random from all the images in the database that do not contain the subject. The advantage of the DPM is that it can correct for small translation and scale mis-alignments of the regions during training. The DPM is then applied to all images on vases that contain the subject's name.

6 Results

For each of the visually consistent clusters obtained in section 3 a strong detector is built through the methods of sections 4 & 5. The total number of DPMs trained is 118, one for

each visually consistent cluster (106 for the gods in table 1 plus 12 for animals).

In order to gauge the success of the strong detectors all the images visually consistent with 'Zeus Seated' and 'Athena Device' across vases containing the respective god's name are manually labeled and annotated to provide a "ground truth" of known positive examples. A detection is deemed correct if it has at least a 0.5 overlap ratio with the ground truth annotation. Figure 6 shows the precision-recall curve for both the DPM and the LDA detectors (trained with the same positive examples). A HOG visualization for the detectors is given in figure 7, and figure 5 shows high scoring detections from the DPM models.

Quantitative Results. The PR curves of figure 6 indicate that the detectors are able to find a large proportion of instances of the subject before a drop in precision. The method succeeds despite large distortions (rounding) of vase decorations, e.g. significant foreshortening. On examining the results, we note that false detections are rarely made on images where the subject is present. Failures tend to occur on images where there is confusion with another figure that is visually consistent with the subject of the cluster, for example a figure sitting (for Zeus) or holding a shield (for Athena).

The recall does not reach one for several reasons: some images are of vase fragments, causing the subject to be disjointed; other images are exposed to lens flare, obscuring large portions of the subject; in a handful of instances the intra-class variation is too extreme for the model to succeed, such as when Zeus is leaning backwards on a chair.

If various components of the algorithm are not used then, as would be expected, the performance is reduced. For Zeus the AP for the full algorithm is 0.5919. When the search space is not reduced before finding windows the AP falls to 0.4331. When no additional positives are sought before training the DPM the AP becomes 0.4657. The most significant drop is when the candidate windows aren't refined from 20 to 4 based on visual consistency (AP: 0.2672); this is due to the presence of false positive windows. A similar loss in performance is observed for Athena.

Qualitative Results. By looking at the top detections for many clusters further observations can be made. The detectors work well when the subject is holding large object such as Apollo with a lyre or Dionysos with a drinking vessel, most likely because these are quite distinct from any other figures. Animals are similarly successful for this reason. Conversely, the detector fails when the distinguishing object held by the god is too small, such as a small vine held by Dionysos. Hermes is a particular problem – although he is in many scenes, he is rarely the focal point and other gods feature more prominently than he does. Detectors trained on red-figure images are less successful that their black-figure counterparts because red-figure images have less standardized figures, increasing intra-class variation.

By looking at the top detections for each cluster we estimate that we have correctly annotated around 3,000 god instances and 2,000 animal instances.

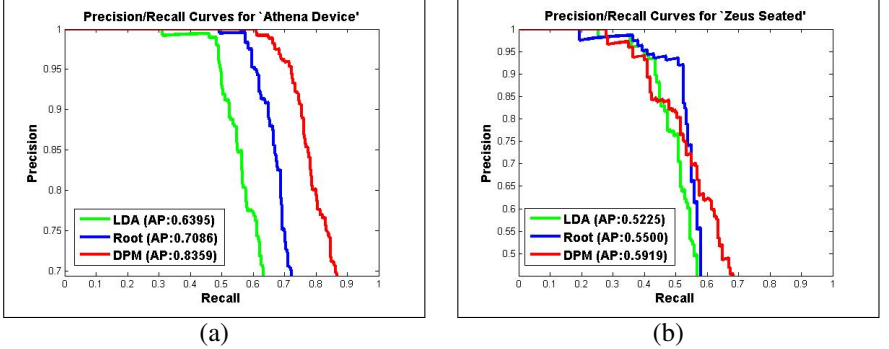
7 Conclusion and Future Work

This paper has provided a working method to solve the "words and pictures" problem of automatically annotating gods and animals on Greek vases. It also has shown that a HOG feature is usable for representing this type of material.

The examples of gods that have been mined are a very useful resource for archaeologists studying classical art, as assembling this type of material (e.g. all depictions of Zeus, aligned and size normalized) manually would take quite some time.

Apart from being a very useful example computer vision application, the method is applicable to other such art/archaeological collections and, more generally, the idea of progressive reduction of visual search space is useful for 'words and pictures' datasets.

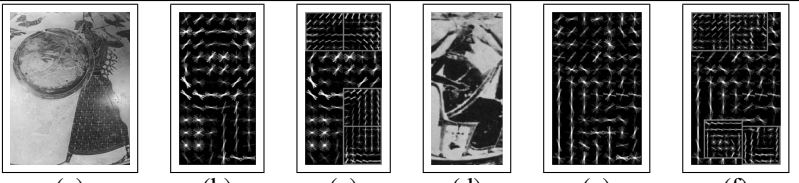
Figure 5: **Top Detections**. Above: Athena Device, Below: Zeus Seated. Note the intra-class variation within each pose .



(a) (b)

Figure 6: **Precision/Recall curves for (a) 'Athena Device' and (b) 'Zeus Seated'**. The green curves are for LDA models, the blue curves are for the DPM root-filters and the red curves are for the full DPM. There are approximately 400 images in the database containing (a) and 200 containing (b).

Future Work will consist of taking fuller advantage of the scene descriptions such as relative positions and pairings of gods. For example, Artemis isn't very distinct but she is often to one side of Apollo ("Apollo playing Kithara between Leto and Artemis"). In addition we will search for the objects the gods possess (Poseidon's Trident or Hermes' winged sandals and winged cap) as these are often quite discriminative.



(a) (b) (c) (d) (e) (f)

Figure 7: **Model Visualizations**. (b) & (e) are the root-filters of the 'Athena Device' DPM and 'Zeus Seated' DPM respectively. (c) & (f) show the configuration of the parts. (a) & (d) are images of the subjects given for comparison.

Acknowledgements. Funding for this research is provided by the EPSRC and ERC grant VisRec no. 228180.

References

[1] The Beazley Archive of Classical Art Pottery Database, Jul 2013. URL http://www.beazley.ox.ac.uk/pottery.

[2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, 1993.

[3] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. IEEE*, 2001.

[4] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *J. Machine Learning Research*, 3:1107–1135, Feb 2003.

[5] T. Berg, A. Berg, J. Edwards, and D. Forsyth. Who's in the Picture. In *NIPS*, 2004.

[6] T. Berg, A. Berg, J. Edwards, M. Mair, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and Faces in the News. In *Proc. CVPR*, 2004.

[7] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *Proc. CVPR*, 2007.

[8] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *Proc. ECCV*, 2006.

[9] N. Dalal and B Triggs. Histogram of Oriented Gradients for Human Detection. In *Proc. CVPR*, volume 2, pages 886–893, 2005.

[10] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012.

[11] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. ECCV*, 2002.

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. CVPR*, pages 1778–1785, 2009.

[13] P. Felzenszwalb, R. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2010.

[14] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 71(3):273–303, 2007.

[15] D. Guo, X.and Liu, B. Jou, M. Zhu, A. Cai, and S. Chang. Robust object co-detection. In *Proc. CVPR*, 2013.

[16] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proc. ECCV*, 2008.

[17] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Proc. ECCV*, 2012.

[18] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[19] L. Jie, B. Caputo, and V. Ferrari. Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. In *Proc. NIPS*, 2009.

[20] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Proc. CVPR*, 2013.

[21] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Prentice Hall, 2000.

[22] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *Proc. ECCV*, 2010.

[23] D. Kurtz, J. Shotton, F. Schroff, Y. Wilks, G. Parker, G. Klyne, and A. Zisserman. CLAROS - Bringing Classical Art to a Global Public, 2009.

[24] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *Proc. ICCV*, 2011.

[25] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[26] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Proc. NIPS*, pages 570–576. MIT Press, 1998.

[27] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proc. 15th International Conf. on Machine Learning*, pages 341–349, 1998.

[28] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *Proc. ACM SIGGRAPH*, 2004.

[29] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proc. ECCV*, 2012.

[30] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, volume 2, pages 1470–1477, 2003.

[31] J. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *Proc. ICCV*, pages 756–763, 2005.